

DOT/FAA/AM-02/12

Office of Aerospace Medicine
Washington, DC 20591

Development of an FAA- EUROCONTROL Technique for the Analysis of Human Error in ATM

Julia Pounds
Civil Aerospace Medical Institute
Federal Aviation Administration
Oklahoma City, OK 73125

Anne Isaac
Human Factors and Manpower Unit
EUROCONTROL
Brussels, Belgium

July 2002

Final Report

DISTRIBUTION STATEMENT A:
Approved for Public Release -
Distribution Unlimited

This document is available to the public
through the National Technical Information
Service, Springfield, VA 22161.



U.S. Department
of Transportation
**Federal Aviation
Administration**

20020910 005

N O T I C E

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents thereof.

Technical Report Documentation Page

1. Report No. DOT/FAA/AM-02/12		2. Government Accession No.		3. Recipient's Catalog No.	
4. Title and Subtitle Development of an FAA-EUROCONTROL Technique for the Analysis of Human Error in ATM				5. Report Date July 2002	
				6. Performing Organization Code	
7. Author(s) Pounds, J., and Isaac, A.				8. Performing Organization Report No.	
9. Performing Organization Name and Address FAA Civil Aerospace Medical Institute P.O. Box 25082 Oklahoma City, OK 73125				10. Work Unit No. (TRAIS)	
				11. Contract or Grant No.	
12. Sponsoring Agency name and Address Office of Aerospace Medicine Federal Aviation Administration 800 Independence Ave, S.W. Washington, DC 20591				13. Type of Report and Period Covered	
				14. Sponsoring Agency Code	
15. Supplemental Notes This work was performed under Task HRR518.					
16. Abstract Human error has been identified as a dominant risk factor in safety-oriented industries such as air traffic control (ATC). However, little is known about the factors leading to human errors in current air traffic management (ATM) systems. The first step toward prevention of human error is to develop an understanding of where it occurs in existing systems and of the system variables which contribute to its occurrence. This paper reports on the project to harmonize the Human Factors Analysis and Classification System (HFACS) and the Human Error Reduction in ATM (HERA) technique. Two groups of air traffic control subject-matter experts (SMEs) participated. The first group analyzed incident cases using each technique and identified the useful concepts from each technique for these cases. The second group evaluated the concepts identified by the first group. Based on these activities, the techniques were deemed to be compatible and harmonization proceeded. Elements from both techniques were retained and many were elaborated based on the SMEs' feedback. The integrated approach, called JANUS, is currently undergoing beta testing by seven European nations and the US Federal Aviation Administration.					
17. Key Words Human Error, Operational Error, Air Traffic Control, Human Performance, Safety				18. Distribution Statement Document is available to the public through the National Technical Information Service; Springfield, Virginia 22161	
19. Security Classif. (of this report) Unclassified		20. Security Classif. (of this page) Unclassified		21. No. of Pages 26	
				22. Price	

Form DOT F 1700.7 (8-72)

Reproduction of completed page authorized

ACKNOWLEDGMENTS

This report summarizes the first phase of work conducted as the part of the FAA/EUROCONTROL Action Plan 12: Management and Reduction of Human Error in ATM. Dr. Manfred Barbarino, EUROCONTROL Programme Manager for the Human Factors and Manpower Unit, and Dr. Mark Rodgers, FAA Chief Scientist and Technical Advisor for Human Factors (AAR-100), were the signatories to the joint agreement. Dr. Paul Krois is the FAA technical point of contact in AAR-100. Dr. Barry Kirwan participated in early stages of the project. All have provided valuable guidance as the research proceeded.

The authors wish to specifically acknowledge the air traffic management subject-matter experts who participated in this research. They were gracious in their willingness and unwavering in their involvement. Without their support and insight, we could not have accomplished this research.

The FAA was represented during the harmonization by John D. Canoles (Manager of Air Traffic Evaluations and Investigations Staff), Anthony Ferrante (Manager of Air Traffic Investigations Division), Tom Carmody (Manager of Air Traffic Evaluations Division), Ralph Abney (FAA-retired), and Don Ellis (FAA-retired). Numerous other FAA air traffic controllers also participated in earlier foundational work leading to this project.

EUROCONTROL was represented by Christer Boll (Swedish Civil Aviation Administration), Rien Doorn (ATC, the Netherlands), Martin Polman (ATC, the Netherlands), Franca Pavlicevic (EUROCONTROL Brussels), and Greg Hindson (ATC, France).

Without them all, we could not have reached this point in the research.

DEVELOPMENT OF AN FAA-EUROCONTROL TECHNIQUE FOR THE ANALYSIS OF HUMAN ERROR IN ATM

INTRODUCTION

Human error has been identified as a dominant risk factor in safety-oriented industries such as air traffic control (ATC). As the capacity and complexity of airspace continue to increase and as ATC develops more advanced interfaces and computerized support technology, the importance of identifying the human factors leading to human error will increase, straining the ability of traditional design practices, alone, to effectively mitigate human error. Therefore, appropriate methods for developing error-tolerant systems are needed. However, little is known about the causal factors leading to human errors in current systems. Thus the first step toward prevention is to develop an understanding of where human error occurs in existing systems.

This paper reports on the project to harmonize two methods for investigating the human factors behind human errors in air traffic safety systems. The *Human Factors Analysis and Classification System* (HFACS) is a human factors taxonomy originally developed for the US Navy to investigate military aviation accidents and is currently being used by the US Federal Aviation Administration (FAA) to investigate civil aviation accidents and causal factors in ATC operational errors (OEs). The *Human Error Reduction in ATM* (Air Traffic Management; HERA) technique is a method of human error identification developed by EUROCONTROL for the retrospective diagnosis of airspace incidents and for prospective analysis during ATM system development.

Activities undertaken to explore the possibility of harmonization depended on input from two groups of air traffic control subject matter experts (SMEs). The first group analyzed incident cases using each technique and identified the useful elements from each technique for these cases. The second group evaluated the elements identified by the first group. Based on these activities, harmonization proceeded, and the techniques were deemed to be compatible. Elements from both techniques were retained and many were elaborated based on the SMEs' feedback. The integrated approach, called JANUS, is currently undergoing beta testing by seven European nations and the FAA.

BACKGROUND

Human Error in Air Traffic Management

Human errors in ATM/ATC have been defined by Isaac and Ruitenberg (1999, pg. 11) as "intended actions which are not correctly executed." Further, Hollnagel, Cacciabue, and Hoc (1995) pointed out that the term, *human error*, can denote a cause, as well as an action. Dekker (1999) proposed that we must go beyond the current, popular models of safety that categorize human errors, viewing human errors as human shortcomings, use concepts such as "loss of situation awareness" as explanations for error, and look to divert blame from the individual to higher levels of the organization. He argued that safety can be better understood by appreciation for the patterns of failure resulting from the affect of limited resources and multiple competing goals. Thus, to comprehensively examine human error in air traffic control, one should consider the possibility of cognitive failure, which may result in an incorrectly executed action, but only as part of a larger matrix of potential failure points.

Past research has demonstrated that breakdowns in cognitive processing such as attention and communication, have contributed to reported operational errors (OEs) in US airspace. An OE is defined as an occurrence attributable to an element of the air traffic system in which: 1) less than the applicable separation minima results between two or more aircraft, or between an aircraft and terrain or obstacles (e.g., operations below minimum vectoring altitude (MVA); equipment/personnel on runways), as required by FAA Order 7110.65 or other national directive, or 2) an aircraft lands or departs on a runway closed to aircraft operations after receiving air traffic authorization, or 3) an aircraft lands or departs on a runway closed to aircraft operations, and it was determined that a Notice To Airmen (NOTAM) regarding the runway closure was not issued to the pilot as required, at an uncontrolled airport (FAA Order 7210.56, 2001). Early analyses by Kinney, Spahn, and Amato (1977) found that 95% of separation violations in en route centers that were classified as operational errors involved errors in attention, judgment, or communications. These same error types have repeatedly been

found in other studies of air traffic control operational errors (e.g., Redding, 1992; Rodgers & Nye, 1993; Schroeder, 1982; Schroeder & Nye, 1993; Stager & Hameluck, 1990).

Analysis of Human Error in ATM

The FAA has several model-based research programs related to identifying and reducing human error in aviation. One of these is the work currently underway at the Civil Aerospace Medical Institute (CAMI) to adapt a previously developed method, Human Factors Analysis and Classification System (HFACS), to the ATC environment for research on human factors related to OEs. EUROCONTROL has also recognized the need for a model-based approach to understanding human error and is pursuing similar work in the Human Error Reduction in ATM (HERA) project (EATMP, 1999a).

Although there are parallels between the FAA and EUROCONTROL objectives regarding to human error, there are differences in the ways the issue of human error is being addressed. For example, both the FAA and EUROCONTROL have focused on human error, cognitive processes, and other operational factors. However, the two techniques vary in distinctive ways. The following sections compare the two techniques on several dimensions: original purpose, theoretical basis, range of concepts covered, data used for analysis, the analysis process, reliability and validation, and the output data. These comparisons between the two techniques are summarized in the Appendix.

The EUROCONTROL Approach - The Human Error Reduction for ATM (HERA) Technique

The development of HERA occurred over the course of six specific research activities.

- Literature from the relevant academic and industrial research findings on human performance models and taxonomies of human error over the past five decades were reviewed.

- Using the results from this review, the most appropriate model of human performance upon which to base the conceptual framework was chosen. The framework selected was the information processing model from the work of Martiniuk (1976) and Wickens (1992). Information processing has proven to be one of the more useful psychological models of performance in various industries. With its emphasis on input, thought, output and feedback, it was judged to be most useful for explaining behavior and also informative for more prac-

tical considerations such as designing new displays, etc. The human information processing model encompasses all relevant ATM behaviors and also allows a focus on certain ATM-specific aspects, such as "the picture" and situation awareness. Thus, it was expected to be a good candidate for a platform upon which to base The Human Error Reduction for ATM (HERA) Technique, if suitably adapted to ATM.

- A review of current and future ATM systems was then undertaken (EATMP, 1999b), as well as a systematic task analysis of controller activities in the tower, terminal, en-route, and oceanic areas. Information from interviews with controllers representing these functional areas was also used to develop not only the HERA technique and taxonomies but also a contextual approach to support the understanding of "how" and "why" controller errors occur. This work also considered future ATM systems such as computerized conflict detection support tools, electronic strips, final approach spacing tools, and data-link technology.

- The chosen conceptual model and framework were then adapted to the ATM context. During this stage, parallel research investigating the controller's mental picture of the traffic was incorporated into the final model, which focused more attention on the role of "working memory" than previous models had done. Performance Shaping Factors (PSFs; e.g., traffic mix, airspace characteristics, procedures, training, equipment, and personal factors) are additional factors relating to error causes, and were thus included in the method. The HERA concept of information processing is shown in Figure 1.

- HERA incorporated some organizational causes into its structure. However, because safety culture was still a developing field, particularly in ATM, elaboration of these elements was left for later phases of HERA, when the safety culture field, itself, could offer more practical guidance on what factors should be included and how they would interplay with the rest of the HERA model.

- Finally, a prototype technique was identified that satisfied the identified criteria for the HERA approach. This technique, which incorporated flow-charts similar to a fault tree method to identify the psychological underpinning of the erroneous behavior, represented the basis for the HERA system.

Thus, HERA places the air traffic incident in its ATM context by identifying the ATC behavior, the equipment used, and the ATC function being performed. Detailed analysis of information processing is set in the context of the controller's working environment and organizational influences.

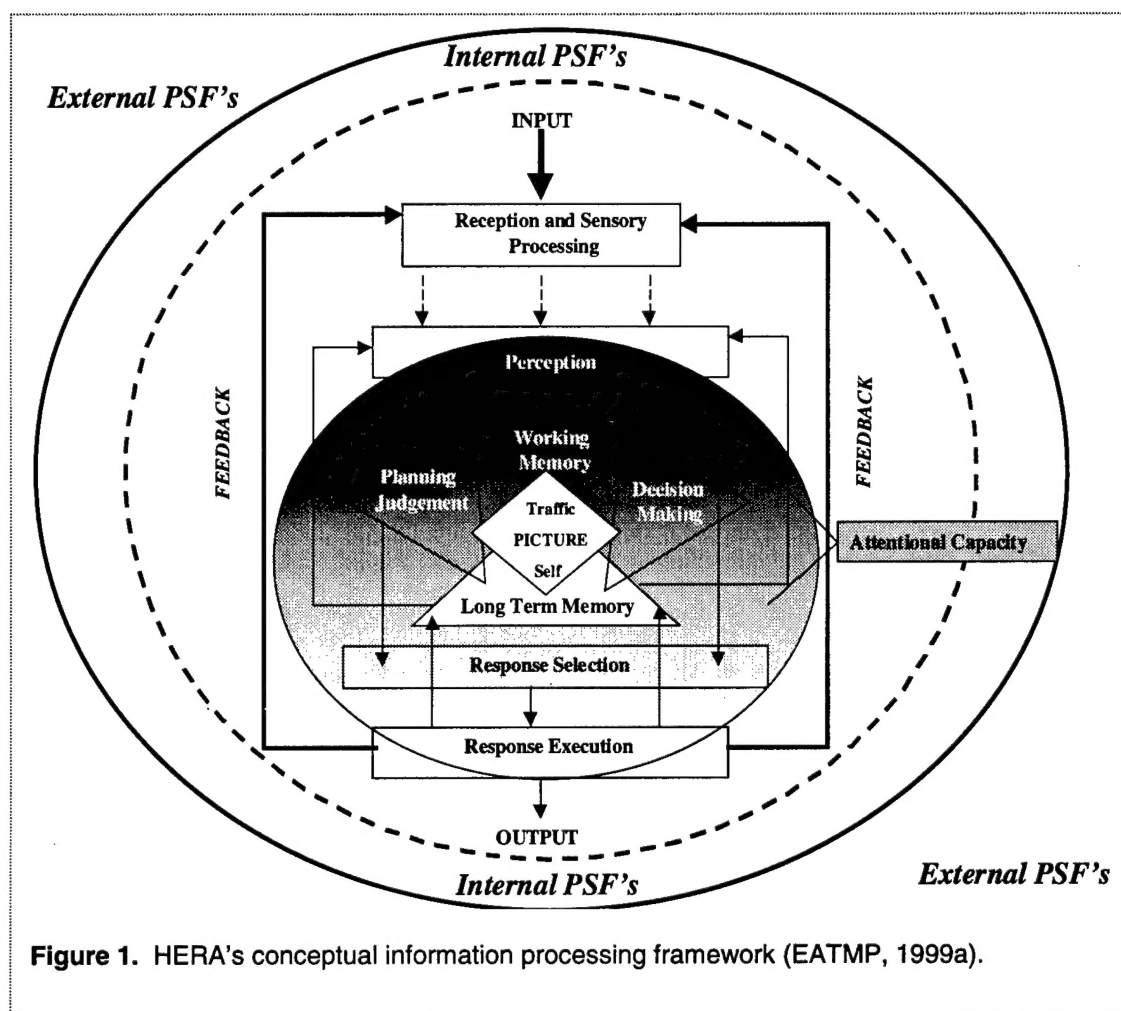


Figure 1. HERA's conceptual information processing framework (EATMP, 1999a).

The FAA Approach - The Human Factors Analysis and Classification System (HFACS)

HFACS (Shappell & Wiegmann, 2000) evolved from the Taxonomy of Unsafe Operations (TOU), a human error approach to aviation accident investigation developed by Shappell and Wiegmann (1997).

- TOU was designed to be a cause-oriented, rather than an outcome-oriented investigation scheme that could be used in multiple occupational settings other than to aviation. The original TOU taxonomy linked accident investigation methods to theory by providing a framework for conducting investigations.

- According to the developers, the taxonomy also provided field accident investigators with a "user-friendly" method for human factors analysis of the accident event. By focusing on underlying causes, rather than only on the failure itself, the method identified the areas requiring interventions. For example, an unsafe act that results from a memory lapse would probably require different interventions than an unsafe act that results from the

operator's willful violation of a rule. For example, TOU analyses by the US Naval Safety Center demonstrated that interventions to avoid controlled flight into terrain should focus on pilots' decision processes. However, TOU did not address other potentially relevant variables such as hardware, software, equipment failures, design flaws, environmental distractions, management, and organizational influences.

- The current version of HFACS used for this project examines instances of human error as part of a complex system. HFACS combines multiple definitions of "human factors" into a coherent taxonomy that includes management and organizational failure points, and adopts a systems approach for investigation and prevention of aviation accidents. Based on the concepts of latent and active failures, HFACS describes four levels of failure: 1) Unsafe Acts, 2) Preconditions for Unsafe Acts, 3) Unsafe Supervision, and 4) Organizational Influences (Shappell & Wiegmann, 2000). The basic HFACS taxonomy is shown in Figure 2.

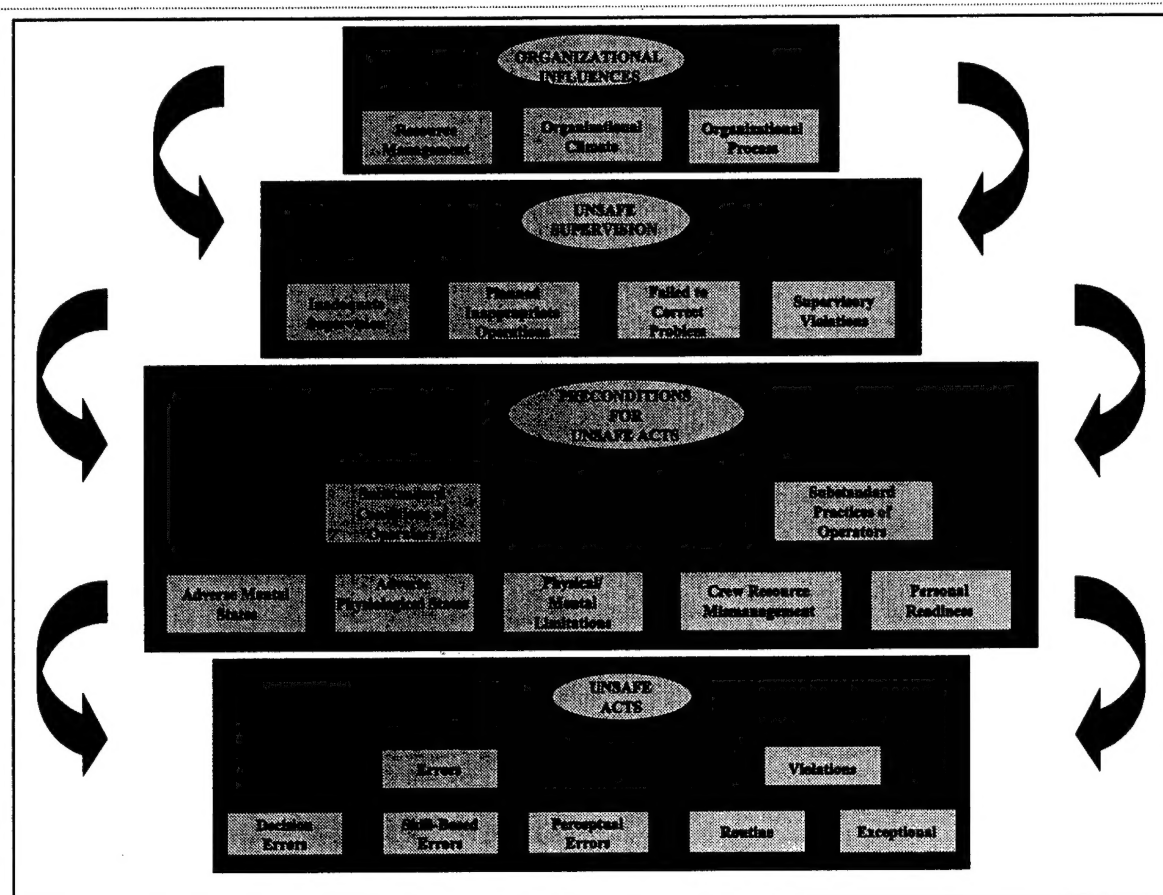


Figure 2. Tiers, categories, and subcategories of the HFACS taxonomy (Shappell & Wiegmann, 2000).

- HFACS was initially developed using aviation mishap data for the purposes of organizing causal factors to identify recurring causal factors and system trends, although the basic taxonomy has other uses. For example, the taxonomy can be adapted to any domain where human error and accidents occur either as a technique to raise awareness for human error and accident prevention or as a causal factors analysis tool (Wiegmann & Shappell, 2001b).

Thus, HFACS, based on Reason's model of unsafe behaviors (1990), places the human error being evaluated as part of a larger systemic problem. To date, the HFACS framework has been used by the US Navy, Marine Corps, Army, Air Force, FAA, US Forest Service, US Coast Guard, and the Canadian Armed Forces.

THE TRANS-ATLANTIC PARTNERSHIP

Comparison of Theoretical Backgrounds

In the development of the Human Error Reduction in ATM technique, a number of different sources were reviewed to determine the concepts that should

be present to adequately represent an ATM model of human error. Several types of modeling approaches were identified: existing human error taxonomies, general psychological models of human performance and error, and methodologies from different industries. The chosen model was embedded in the ATM framework, including current and future ATC tasks and requirements (EATMP, 1999a).

HERA conceived of the operator as part of a larger system. The process took investigation of the human performance factors beyond the individual and included different facets of the situation to try and understand the mechanisms and context leading to the error.

In the initial development of HFACS, several literatures were reviewed, including human factors, industrial safety, and crew resource management. This review revealed that methods for identifying human failures should first make assumptions about the existence of the structure and process of human cognition (Wiegmann & Shappell, 1997). Further, the behavioral acts resulting from the decision and whether

the acts were intentional or unintentional, were also relevant for understanding. Based on this review, three approaches were integrated into one taxonomy: a variant of the four-stage human information processing model (Wickens & Flach, 1988), a model of cognitive malfunction (O'Hare, Wiggins, Batt, & Morrison, 1994; Rasmussen, 1982), and a model of unsafe behaviors (Reason, 1990). Other categories not covered by these frameworks were added to HFACS to include social variables such as teamwork (Jensen, 1995) and physiological variables (e.g., fatigue), supervision, and contextual factors. Thus, HFACS incorporates human, environment, and organization elements as identified by Bird (1974), Edwards (1988), and Heinrich, Petersen, and Roos (1931).

Discussion of Theoretical Backgrounds

Both techniques incorporated common elements from human error models and general psychological theories. Comparison of the two techniques suggests that the HERA technique was formulated after identifying useful frameworks relevant in ATM/ATC. Further, HERA development included models of ATC performance and of current and future ATC task and behavioral requirements. In contrast, the development of HFACS was based on a set of models drawn from psychology, aviation, and accident and industrial human error identification. Although later adaptations have expanded the HFACS model to other domains (e.g., aviation maintenance), the original model was developed for the investigation of US Naval aviation accidents and incidents. Thus, the taxonomy was not originally developed to represent ATC concepts, specifically, although the general concepts captured in the HFACS tiers, categories, and subcategories seemed to be generally applicable to ATC. Both HFACS and HERA include some of the same theoretical concepts but at different levels of granularity. For example, the specificity of HERA's identification of psychological error mechanisms is not captured explicitly in the HFACS technique. The HFACS analyst is, however, required to categorize cognitive processes.

Comparison of Conceptual Coverage

The HERA analysis examines the human error event in relation to contextual factors such as the task engaged in at the time of the error and equipment being used. HERA allows the analyst to explain the context by assigning of appropriate information keywords. The analysis identifies which of the cognitive dimensions of perception and vigilance, working

memory, long-term memory, or judgment, planning, and decision making are being relied upon when the human error occurs.

HERA uses Cognitive Domains (CDs; e.g., sensory reception, perception, working memory) to provide a structure to organize the errors. Each CD is further analyzed in terms of the Internal Error Modes (IEMs) and Psychological Error Mechanism (PEMs) with which it is associated. IEMs represent the internal outcome of an error (e.g., misidentification, late detection, misjudgment) and PEMs describe the psychological mechanism (e.g., perceptual tunneling) associated with the IEMs. Analysis of each error also includes the identification of Performance Shaping Factors (PSFs) (variables like organizational influences, supervision, team and personal issues, and traffic characteristics) and External Error Mode (EEM) — the expression of the error — such as *action performed too late*.

HFACS classifies the error event by examining causal factors leading to the final outcome — the Unsafe Act. Based on the concept of active and latent "failure points" in the system (Reason, 1990), HFACS identifies those "holes" in the system's defenses that could all align to cause a human error. HFACS, conceptually, encourages the analyst to capture both the depth and breadth of the situation.

As noted in a preceding section, these HFACS causal factors are organized into four tiers (Unsafe Acts, Preconditions for Unsafe Acts, Unsafe Supervision, and Organizational Influences). Each tier is subdivided into categories and subcategories. The Unsafe Acts tier captures the active failure — the identified act committed by the operator. This can either be categorized as errors in the operator's decision, skill, or perception; or routine or exceptional rule violations. To understand why the event took place, the action is examined in terms of the precursors — the preconditions for the unsafe act. This tier is subdivided into categories of substandard conditions of operators (i.e., adverse mental and physical states, and mental or physical limitations), and substandard practices of operators (i.e., crew resource mismanagement and personal readiness). The chain of causality is traced backwards to include the possibility of Unsafe Supervision. At this level, HFACS examines the possibility of inadequate supervision, planned inappropriate operations, failure to correct problems, and supervisory violations. Organizational influences and upper level management factors are captured in the Organizational Influences tier, which include categories of resource management, organizational climate, and organizational process.

Discussion of Conceptual Coverage

Both HFACS and HERA view the individual operator as an element in a larger safety system. Conceptually, both techniques analyze the error event by considering the relationships between elements in the system. Both techniques also examine individual errors and the situational and organizational factors surrounding the event. The strength of the HFACS technique is that it forces the analyst to capture the conceptual depth and breadth of the system view by moving from the individual act to the preconditions, supervision, and organizational influences. HERA's strength is that it provides a fine-grained analysis of the individual's cognitive processes to identify those that led to the error event. Thus, the conceptual similarities and differences between the techniques are not so much related to which concepts are included but rather, the differences between where the primary analytic effort is invested.

Comparison of Analytic Methods

Analysts using the HERA taxonomy work from the narrative description of an airspace event (incident) that results from the incident investigation to identify each reported human error. Each human error is analyzed consecutively as a separate unit of analysis using the HERA technique. The analyst is advised against making speculative assumptions and any assumptions, by the analyst are noted on the form; clear "stop rules" are followed in this regard.

A brief but specific description of each individual human error point is entered into the HERA analysis form, including how the error was detected and recovered, if known. The type of effect the error had on the following events (Causal, Contributory, Compounding, Non-Contributory) is also recorded. Contextual factors associated with the error (Task, Equipment, and Information) are identified from checklists available for the analyst's reference. After the Error or Violation types are identified, the analyst is led through a series of flowcharts to identify the Cognitive Domains and resultant Internal Error Mode and Psychological Error Mechanism level. Any Performance Shaping Factor (i.e., traffic, procedures, training, teamwork, HMI, personal factors, and organizational factors) that might have prompted the error or made its occurrence more likely is also recorded.

Coders employing HFACS retrospectively for aviation mishaps, for example, use the list of causal factors identified by the original incident investigators from their on-site reports. Incident narratives are used to provide further context, but analysts using HFACS evaluate only the reported information provided in

accident reports and accompanying narrative materials and are encouraged to resist making assumptions about information that is not specifically reported. Analysts first identify whether an unsafe act was committed by the operator. If so, the analyst classifies the act as an Error or a willful Violation of the rules. Then, the analyst classifies the error according to the appropriate subcategory. Proceeding from the Unsafe Act, causal factors related to the error are identified and arranged in a sequence, moving backwards from the time of the event. Each causal factor is classified according to the tiers, categories, and subcategories of the HFACS taxonomy.

Discussion of Analytic Methods

The two techniques differ in several ways. One is in the identification of the causal factors used for analysis. HERA analysts work primarily from the incident narrative and use the HERA process to identify and classify causal factors. HFACS coders, as reported in the research using aviation mishap reports, work primarily with a list of causal factors which identified by incident investigators. These causal factors are classified by expert coders into the appropriate category of the HFACS taxonomy, using the incident narrative description to provide context and clarification.

The techniques appear to also differ in the unit of analysis adopted. In the conduct of the analyses, the HERA analysis is performed on each identified human error point as the unit of analysis, whereas HFACS analysis is performed using the incident as the unit of analysis. To accomplish this, HERA proceeds from the beginning of the description and moves forward in time. HFACS analysis begins at the terminal event and proceeds backward in time.

Because of these differences, the two techniques result in somewhat different types of output data. The HERA process permits analysts to elaborate on the results of the original investigation. HERA analysis of an airspace incident can result in a set of categories, psychological mechanisms, and performance-shaping factors for each human error analyzed in the incident as it unfolded over time. For example, HERA analysis of one incident can result in data from multiple human errors, multiple types of tasks, etc. For the analyst, this creates a description of how human errors cascade and propagate through time and result in an incident.

HFACS analysis results in categorization of causal factors surrounding the Unsafe Act most proximally associated with the final outcome of the incident; linkages between variables are made and preserved as relational databases. Similar to the HERA process,

HFACS analysts are encouraged *not* to “reinvestigate” the accident and to resist making inferences about the incident beyond the information given. Working backward from the final Unsafe Act, the analyst classifies any associated events that contributed to the final outcome. The output of HFACS analysis consists of two types of data sets. One is the frequency of occurrence for the causal factor subcategories, categories, and tiers. These can be collapsed and reported as summary data to capture trends at the desired level of analysis, e.g., Decision Errors, Unsafe Acts, and Unsafe Supervision. For the second data set, HFACS analysts also record a description of each causal factor coded.

To perform retrospective analysis of an incident, both techniques rely on summary data from other investigators. In both cases the analysts using these techniques are urged to resist making their own assumptions beyond the process and data given. Both techniques place human error in a situational context.

Comparison of the Reliabilities

Any useful human error framework should be broadly applicable. That is, different users analyzing the same event using the method should identify similar factors. With this goal in mind, both HERA and HFACS were developed in an iterative manner. Both were developed against the criteria of Cohen's Kappa, an index of agreement between multiple coders, corrected for chance. Values of $k = .40$ or less are considered “poor” agreement, while values of $k = .75$ or greater are considered “excellent” levels of agreement (Fleiss, 1981).

A large validity and reliability study was conducted to test the HERA technique for consistency across users and across reports originating from different nations (EATMP, 2000). A total of 26 people participated in these validation trials. One concern was whether a classification system with a large number of categories could be used reliably by different groups of users. Thus the validation also compared the coders' professional background and length of training using the method. All studies used incident reports from European airspace events. For agreement on the general category of Cognitive Domain, Kappa values ranged from .44 to .50. Analysis by job function of the coders (ATM, human factors researchers, and incident investigators) demonstrated that the primary target group of users (incident investigators) showed the highest agreement overall (Kappa .61). ATM and the overall Kappas for researchers' agreement were .23 and .43, respectively. Agreement between coders declined as the level of analysis became finer-grained and psychological specificity of the technique increased.

Several reliability studies have been conducted during the development of the HFACS model. Initial studies were conducted using lists of causal factors from US Navy, Marine Corps, and Air Force aviation mishap reports (Shappell & Wiegmann, 1997). In a summary of five studies conducted during the development of the model, each using three independent raters, Wiegmann and Shappell (2001a) reported that inter-coder reliability between rater pairs ranged from .60 to .95, using pilots and aviation psychologists as coders. Causal factor reports of commercial aviation accidents from the US National Transportation Safety Board and the FAA coded by a commercially rated pilot and an aviation psychologist resulted in a Kappa of between .65 and .75 (Wiegmann & Shappell, 2001a) and .72 when general aviation accidents were independently coded by five pilots (Shappell & Wiegmann, 2001).

Discussion of Reliabilities

Both methods had what their developers considered to be successful validation trials, although levels of Kappa varied widely between the methods. Although initial comparison of the variation in Kappa values suggests that the two techniques have important differences in how reliable they would be for applied users, the techniques and processes for using the techniques also have important differences that may have influenced the measure of inter-rater agreement. Thus to compare Kappas between the two techniques may be comparing apples to oranges.

One difference is the information about the incident used as input for the analysis. HERA uses the ATC incident investigation narrative report written by the incident investigators as the primary input. HFACS uses as primary input a list of causal factors identified by incident investigators. Thus the quality and differences in the inputs may influence the range of inter-analyst agreement.

Another difference is the decision patterns required of coders involved in the development studies. The primary role of an analyst using HERA is to make “Yes-No” judgments based on the incident narrative in response to a series of predetermined and contingent queries, which are diagnostic about the individual's cognitive processes, and then also to identify related performance shaping causal factors from checklists. The queries lead the analyst through three levels of questions in flowchart formats, therefore requiring multiple decision points for identification of each element. In contrast, the diagnostic decisions required to identify the list of causal factors used for the HFACS method have already been

performed by the initial incident investigators. Therefore, the primary role of an HFACS analyst is to decide where each causal factor from this list of pre-determined items should be classified into the taxonomy.

Based on these differences, it is not clear what added information a comparison of reliabilities between the two models would provide to the harmonization decision. After a harmonized technique is developed, its reliability will have to be assessed on its own terms.

THE HARMONIZATION PROCESS

The harmonization process to create a technique using the strengths of both techniques was undertaken in three separate but associated phases. Phase 1 analyses compared techniques and developed materials for Phase 2 analyses. In Phase 2, operational personnel provided their opinions about the relative utility of concepts from each technique. Phase 3 produced the harmonized technique. The principle investigators from the FAA and EUROCONTROL organized and led all three phases of the harmonization.

Phase 1 – Comparing the Two Techniques

The above comparisons between the two techniques revealed similarities and differences. The goal of the first phase was to select those concepts within each technique that would be most useful to describe OEs.

The main activity was to have several subject matter experts (SMEs) analyze incident reports using both techniques and to agree with about each analysis and its output. The output from the analyses was used in two ways. First, the output from each technique was examined for similarities and differences in the concepts used from each technique. Second, the output was used in the second phase of activity which was structured to identify which concepts were useful to operational incident investigators.

Participants

Air traffic control SMEs with experience in operational incident investigations were recruited to participate in the analysis of the different techniques: two representatives from EUROCONTROL and two from the US. The two European SMEs were incident investigators. The two US SMEs were retired FAA air traffic control specialists (ATCSs) who were teaching management and quality assurance courses at the FAA Academy in Oklahoma City. Three human factors researchers also participated: the co-principal investigator from EUROCONTROL, who was familiar

with European incident investigations, the FAA's co-principle investigator, and the FAA's ATC human factors program manager. Both of the latter were familiar with the US incident reporting process.

SME Training

The four ATC subject matter experts, two representing EUROCONTROL and two representing the FAA, were familiarized with the concepts of each approach and performed the analyses for Phase 1. The EUROCONTROL SMEs were already familiar with the HERA technique, having participated in HERA development activities. The FAA SMEs were familiar with both approaches, having participated in previous consensus analysis of 50 operational error narratives and discussion activities to adapt HFACS and HERA for use in the US FAA ATC environments, although the original HFACS taxonomy and HERA technique were used for this and all subsequent activities.

Prior to meeting for the consensus analysis, materials and background information for each technique were exchanged, including sample incident narratives from Europe and the US. The EUROCONTROL SMEs also participated with FAA personnel in a four-hour discussion of the HFACS taxonomy via video teleconference. During this session, participants discussed the definitions of each classification as they might relate to ATC behaviors and used the definitions to identify causal factors in two sample US incident report narratives.

Naturally, more intensive training would be required before analysts would consider themselves proficient in the techniques for the purpose of in-depth identification and analysis of incident causal factors. The goals of these Action Plan 12 activities were less encompassing. They were to employ the SMEs to, first, determine whether harmonization was feasible or should even be attempted and, if so, determine which parts of HFACS and HERA should be retained as part of the harmonized technique and which should not. Thus, the research relied on their ATC expertise, their relative, rather than absolute, familiarity with both techniques, a balanced analytical approach using within-subjects designs, and a representative sample of incident narratives from Europe and the US.

Materials and Procedures

Twenty incident cases (10 European and 10 US) were selected to represent different types of possible scenarios, e.g., from terminal and en route. Not surprisingly, the formats were different but allowed the analyses to be undertaken.

The recording form for HERA had eight sections (i.e., Task, Equipment, Information, Error Type, CDs, IEMs, PEMs, and PSFs). The data recording template for HFACS had three classification categories (i.e., Tier, Category, Sub-category).

Differences between HERA and HFACS were compared by examining the levels and concepts from the two techniques. Together, 455 concepts (terms) from HERA and HFACS were used that were potentially important for human factors analysis — 414 from HERA (91%) and 41 (9%) from HFACS. These were distributed within each technique as shown in Table 1. The HFACS Categories (C; e.g., Errors), Subcategories (S; e.g., Decision Errors), and types (e.g., procedural decision errors) within each Tier (T) are not listed here but are rolled into the Tiers and represented in the numbers shown. The Tiers, Categories, and Subcategories are shown in Figure 2.

All participants were first familiarized with both the HERA and HFACS techniques, as described in the training for Phase 1, and worked independently to analyze each of the 20 incident cases prior to convening at the Civil Aerospace Medical Institute in Oklahoma City for a joint meeting. At the meeting, participants spent three days in a group session comparing their results from the individual analyses for 10 of the 20 cases (5 European and 5 US). The goal was to reach agreement on the error points, analysis, and resultant causal factors identified for each incident. The HERA and HFACS causal factors were discussed for each incident and any disagreements were resolved by the SMEs so that a list of error points and elements from each technique for each incident case could be obtained in preparation for Phase 2.

Results

Output from the analysis of each incident was a list of items (identified error events) and the associated human factors terms resulting from the SMEs' analyses. An illustration of the output from the analysis of one error item is shown in the boxed portion of Table 2. Terms 1-9 are the concepts resulting from the SMEs' HERA analysis; terms 10-13 are the concepts output from their HFACS analysis.

To understand the relative contribution of each method, the terms generated in Phase 1 from the SME consensus analysis of all ten cases were compiled into one list. Many of the terms had been identified in more than one case analysis. Overall, the resulting list contained 1818 data points representing the terms used: 1156 (63.6%) from the HERA analyses and 662 (36.4%) from the HFACS analyses.

Because the relative conceptual contribution of each method to the consensus analysis was the primary interest here, duplicate items were removed from the data to eliminate double counting. This resulted in a list of 126 unique concepts: 98 (77.8%) from HERA and 28 (22.2%) from HFACS.

Nevertheless, although the percentage of HERA concepts relative to the HFACS concepts from these analyses differed from the initial availability of 91% and 9%, respectively, the results from Phase 1 did not provide a final answer and only revealed that both techniques contained useful elements upon which the harmonized technique could be built. To gain further understanding about which elements should be retained from each technique and which should be eliminated, the following Phase 2 analysis was conducted.

Table 1. Relative number of concepts contributing to the analysis

HERA Sections	HFACS Tiers
28 - Tasks	13 - Unsafe Acts
85 - Equipment and Information items	9 - Preconditions
27 - External Error /Violation types	5 - Unsafe Supervision
4 - Cognitive Domains with Internal Error	14 - Organizational Influences
Modes encompassing 67 concepts :	
24 - Perception and Vigilance	
17 - Memory	
15 - Planning, & Decision Making	
11 - Response Execution	
207 - Performance Shaping Factors	

Table 2. Example of results from HERA and HFACS Analyses.

	Technique		Phase 2 Analysis	
	HERA	HFACS	Mean Rank	Mean Score
Incident: 11, Situation: Arrival a/c was descended to an altitude that put it in conflict with an overflight a/c.				
Item 1: Controller missed an incorrect altitude readback.				
1. R/T Communications -- read-back	Task		5	.05
2. Descent	Keyword		10.2	.11
3. Clearance	Keyword		8.2	.09
4. Altitude	Keyword		9.2	.10
5. Incorrect information received/recorded	External Error Mode		8.8	.10
6. Perception and Vigilance	Cognitive Domain		4.6	.05
7. Hearback/No Detection-auditory	Internal Error Mode		5.8	.06
8. Expectation bias	Psychological Error Mechanism		6.6	.07
9. Pilot breach of R/T Standards	Performance Shaping Factor		10	.11
10. Skill-based error		T1, C1, S2	6	.07
11. Attention error		T1, C1, S2, Failure	2.2	.02
12. Error		T1, C1	6.4	.07
13. Unsafe act		T1	8	.09

Note. HFACS Levels: T = Tier, C = Category, S = Subcategory. N analysts = 5.

Phase 2 – Analyzing the Two Techniques

The purpose of the second phase was to use the output from Phase 1 to a) identify the most useful concepts from each technique for operational error investigations, b) determine the depth of detail that operational personnel found most useful for retrospective analysis of incidents, and c) evaluate the advantages and disadvantages of each technique as an operational tool. To accomplish this, a panel of experts with experience in both operational investigations and the development of associated mitigation strategies was convened, given some familiarity and practice with each technique, asked to rank elements from each technique, and then were asked for feedback on the relative strengths and weaknesses of each technique.

Participants

The Phase 2 analyses meeting was held at the Institute of Air Navigation Services (IANS) in Luxembourg. Three SMEs from Europe and three from the US were chosen to participate in this expert forum because of their operational expertise, knowledge, and experience with investigation of operational incidents.

Prior to the meeting, they had had no experience with either the HERA or HFACS technique. Two SMEs from the CAMI meeting also participated to clarify any questions about the data.

Materials and Procedures

Before completing the ranking task, the participants were divided into teams, each having both US and European experts. They were given general instructions by the researchers and the facilitating SMEs about conducting HERA and HFACS analyses and were then given one European and one US incident narrative to practice each technique. Two of the experts from the Phase 1 meeting monitored the groups to answer any technical questions about the methods. Each team walked through a consensual analysis for each of the two incident cases (one European, one US) using each method. The order of cases and method used were counterbalanced. This activity was designed to provide experience identifying causal factors from narratives and with using both techniques for analysis before they ranked elements of each technique and before they were asked for feedback about overall strengths and weaknesses of the techniques.

Data from the consensual analysis of the ten cases analyzed in Phase 1 by the SME group convened at CAMI were presented as follows. An incident example from the materials is shown in the boxed section of Table 2. For reference, the example also lists the technique and concept that each Term represented. (The participants did not have this information, however.)

It must be noted that only the elements from each approach that were data output from the Phase 1 analyses were carried over to Phase 2, making the results of Phase 2 analyses directly contingent on the sample of incident reports selected for the consensus analysis. Certainly, it is conceivable that a different sample of incident OE reports for the consensus analysis may have produced a somewhat different set of elements to be used in Phase 2. To mitigate this, the concepts from each technique not selected in the Phase 1 consensus analyses ("orphans") were also ranked by the SMEs who participated in Phase 2. They were listed and ranked without any framing situation and they were kept to be analyzed, should either approach dominate the Phase 2 ranking data. The scores from the "orphan" data are not reported here.

To prepare the materials, first, an overall *Incident Situation* statement was generated to summarize the event. In the Table 2 example, the Situation was that the *Arrival altc was descended to an altitude that put it in conflict with an overflight altc*. The critical points (the human errors by ATC) identified and analyzed in Phase 1 were listed as *Items*. In the example, the first Item (the first critical point of human error by ATC) was that the *Controller missed an incorrect altitude readback*. A total of 40 Items were presented to the SMEs for their analysis.

The number of Items within Incident Situations ranged from 2 to 7 (mean = 4, mode = 5). The *Terms* output from the Phase 1 analysis were listed under each Item. In the example, there were 13 Terms listed under this Item. The number of Terms to be ranked within Items over all Situations ranged from 2 to 26 (mean = 9.1, mode = 13). Overall, a total of 363 Terms — 228 Terms from HERA (62.8%) and 135 (37.2%) Terms from HFACS appeared across Incidents for ranking. (In the right two columns of Table 2 are shown the results from the later Phase 2 analysis for each Term for this incident example.)

The members of the expert forum, working individually, ranked the Terms according to how important each would be (relative to the other Terms in the set) in understanding the Incident Situation using the following method: 1 = Most Important to N = Least Important. Because the number of Items under each Incident Situation was not held constant, N, the

upper limit on the range of scale values, was dependant upon the number of other Terms in its list. Each technique was not equally represented in each list, and 11 Items did not have any HERA Terms listed for ranking. These Items had 2-3 HFACS Terms listed, and examination revealed that they listed primarily supervisory and organizational vulnerabilities.

Results

These results are organized and presented for reporting comprehensibility. However, they were developed iteratively over the course of the harmonization activities. The analyses were not conducted to conclusively identify the merits of one technique over the other but to help the researchers identify the path to a harmonized taxonomy.

Utility of Terms

To identify which concepts from each technique were relatively more and less useful to the experts, the rank of each Term was converted to a score that both represented the number of options competing for ranking with it under that Item and could also be compared across Items.¹ For example, if there were three Terms ranked under the Item, then the denominator used to calculate that Term's score was $3+2+1=6$. Similarly, if seven Terms were ranked, the denominator used was 28. The score for the Term was then calculated by dividing the ranking for that Term by the calculated denominator specific to the group of Terms under the specific Item. The resulting scores can range from 0 to 1, with lower scores indicating a higher ranking adjusted for number of possible Terms competing for that ranking. This method could be taken to a finer level of detail if we took into account the relative contribution of each technique to each list of Items, but for the present purposes, this level of analysis was not conducted.

Ranking data from all experts resulted in data for 1,818 Terms — 1,156 Terms from HERA and 662 from HFACS. Of these, 5 HFACS Terms and 11 HERA Terms received no rankings by the experts and were assigned a ranking of 0. Examination of these Terms revealed no systematic pattern of omission associated with either technique. Eight of the 16 omitted rankings occurred in one of the Item lists having 26 Terms and 3 of the omissions from an Item list of 14 Terms. The remaining 5 omissions resulted from one SME overlooking an Item having 5 Terms.

¹ We thank Dr. David J. Weiss from California State University—Los Angeles for this technique.

Table 3 shows the overall scores for each technique. In general, the expert forum rated the HERA items as being relatively more useful descriptors for enabling their understanding of the incidents' causal factors. However, the intent of the project was not to choose one as being better than the other but rather, to determine whether and how to harmonize the two, so the data were examined in more detail.

Relative Utility of Techniques

The scores were then examined in more detail, first within each technique and then between techniques. To maintain some level of comparability, the comparisons were made at similar levels for each technique. That is, HFACS Tier and Category data were compared with data representing HERA's Sections.

The mean scores for each HERA section are shown in Table 4. Scores have been ordered from lowest (Cognitive Domain) to highest (Internal Error Mode). Within the HERA technique, the expert forum showed a relative preference for those items that were descriptive of the Cognitive Domain (e.g., perception) associated with each critical point and Psychological Error Mechanism (e.g., visual search failure), but ranked Internal Error Mode items describing how the error was manifested internally (e.g., no detection—visual) as being less useful.

The same method was used to compare Terms within the HFACS technique at the Tier and Category levels. Table 5 shows the mean scores associated with HFACS Terms at these levels. Note that the

Subcategories has been summarized at the Category level, and not every HFACS tier and category were represented in the data. Some were eliminated from further consideration by the Phase 1 activity. These data are inconclusive as to whether this was due to the narratives selected for the experimental tests or the nature of the concepts.

Examination of the mean scores for HFACS Terms suggest that the expert forum preferred terms describing the controller's behavior (e.g., Unsafe Act mean score was .09). However, the scores suggest that the HFACS terms addressing organizational influences (.30), preconditions (.33), and supervision (.40) were perceived as relatively less useful, compared with information about the individual's Unsafe Acts.

To compare both techniques, the Terms from Tables 4 and 5 were ordered from lowest to highest. As noted earlier, a conceptual level of equivalency between HERA Sections and HFACS Tiers/Categories was presumed. The results of this ordering are shown in Table 6.

Based on these scores, most of the available HERA terms were rated as having relatively more utility for ATC operations than the available HFACS terms, although the highest scoring terms in both techniques related to the individual's behavior. For example, the HERA Psychological Error Mechanisms such as *making an incorrect assumption* were ranked highly. More detailed examination of this category showed that the terms rated the most useful information for incident investigation with the most frequency were: *making an incorrect assumption* (N = 25, score = .03), *expectation*

Table 3. Analysis of Ratings for Each Technique Over All Items

	<u>N Scores</u>	<u>Mean Score</u>	<u>Std. Dev.</u>	<u>Minimum</u>	<u>Maximum</u>
HERA	1156	.08	.06	0	.40
HFACS	662	.16	.16	0	.67

Table 4. Mean Scores for HERA Sections

<u>HERA Section</u>	<u>N Ratings</u>	<u>Mean Score</u>	<u>Std Dev</u>	<u>Min</u>	<u>Max</u>
1. Cognitive Domain	95	0.04	0.03	0.01	0.16
2. Psychological Error Mechanism Level	110	0.06	0.04	0	0.20
3. External Error/Violation Type	130	0.08	0.04	0.01	0.19
4. Task	151	0.08	0.06	0	0.33
5. Performance Shaping Factors	300	0.09	0.07	0	0.40
6. Information and Equipment	265	0.10	0.04	0.01	0.30
7. Internal Error Mode	105	0.54	0.03	0	0.13

Table 5. Mean Scores for HFACS Tiers and Tier Categories

<i>HFACS Term</i>	<u>N</u> <u>Ratings</u>	<u>Mean</u> <u>Score</u>	<u>Std Dev</u>	<u>Min</u>	<u>Max</u>
1. Unsafe Acts	120	.11	.06	0	.29
1.a Errors	335	.07	.05	.002	.22
1.b Violations	20	.21	.11	.05	.40
Overall Tier Emphasis	475	.09	.06	0	.40
2. Preconditions for Unsafe Acts	5	.43	.09	.33	.50
2.a Substandard Conditions of Operators	10	.28	.14	.17	.50
2.b Substandard Practices of Operators	--	--	--	--	--
Overall Tier Emphasis	15	.33	.14	.17	.50
3. Unsafe Supervision	26	.42	.21	.03	.67
3.a Inadequate Supervision	5	.53	.18	.33	.67
3.b Planned Inappropriate Operations	31	.42	.21	.06	.67
3.c Failed to Correct Problem	5	.08	.07	0	.20
3.d Supervisory Violations	--	--	--	--	--
Overall Tier Emphasis	67	.40	.22	0	.67
4. Organizational Influences	35	.31	.14	0	.50
4.a Resource Management	40	.26	.14	0	.50
4.b Organizational Climate	10	.35	.15	.17	.50
4.c Operational Process	20	.36	.12	.17	.50
Overall Tier Emphasis	105	.30	.14	0	.50

Table 6. Ordering of HERA and HFACS Concepts

		<u>N</u> <u>Ratings</u>	<u>Mean</u> <u>Score</u>	<u>St. Dev.</u>	<u>Min.</u>	<u>Max.</u>
HERA	Cognitive Domain	95	0.04	0.03	0.01	0.16
HERA	Psychological Error Mechanisms	110	0.06	0.04	0	0.2
HFACS	Errors	335	0.07	0.05	0.002	0.22
HFACS	Failed to Correct Problem	5	0.08	0.07	0	0.2
HERA	External Error/Violation Type	130	0.08	0.04	0.01	0.19
HERA	Task	151	0.08	0.06	0	0.33
HERA	Performance Shaping Factors	300	0.09	0.07	0	0.4
HERA	Information and Equipment	265	0.1	0.04	0.01	0.3
HFACS	Unsafe Acts	120	0.11	0.06	0	0.29
HFACS	Violations	20	0.21	0.11	0.05	0.4
HFACS	Resource Management	40	0.26	0.14	0	0.5
HFACS	Substandard Conditions of Operators	10	0.28	0.14	0.17	0.5
HFACS	Organizational Influences	35	0.31	0.14	0	0.5
HFACS	Organizational Climate	10	0.35	0.15	0.17	0.5
HFACS	Operational Process	20	0.36	0.12	0.17	0.5
HFACS	Unsafe Supervision	26	0.42	0.21	0.03	0.67
HFACS	Planned Inappropriate Operations	31	0.42	0.21	0.06	0.67
HFACS	Preconditions for Unsafe Acts	5	0.43	0.09	0.33	0.5
HFACS	Inadequate Supervision	5	0.53	0.18	0.33	0.67
HERA	Internal Error Mode	105	0.54	0.03	0	0.13
HFACS	Substandard Practices of Operators	--	--	--	--	--
HFACS	Supervisory Violations	--	--	--	--	--

bias (N = 20, score = .06), *failure to monitor* (N = 15, score = .04), and *failure to integrate information* (N = 15, score = .08). Similarly, the highest ranked HFACS terms also related to individual behavior, including such processes as choice decision errors or attention failure errors. Further examination of this category revealed that the terms rated the most useful information for incident investigation with the most frequency were *skill-based error* (N = 50, score = .09) and *decision error* (N = 105, score = .06), with the skill-based error of *attention failure* rated as a highly informative concept (N = 55, score = .04).

Interpretation and generalization of the results in Table 6 should be done with some care. Remember that the incidents selected for Phase 1 analysis were selected to be representative of air traffic facility types and descriptive enough so as to exercise both techniques. Thus, the selection of incident cases may have included some unknown bias into the resultant set of concepts output from Phase 1 and used in Table 6. Several terms from both techniques were eliminated from this data by the Phase 1 activities. Therefore, the data in Table 6 are representative only of the terms included in Phase 2 activities. The bottom two rows in Table 6 represent two concepts not used during Phase 1 analyses, and so they did not carry over for Phase 2 ranking, possibly as a result of narrative selection.

Similarly, the overall scores shown in Table 3 may have reflected the absolute number of Terms from each technique available for ranking by the SMEs. The weighted scoring algorithm to transform rankings into scores for each Term was used to address this possibility.

Also, current incident reporting processes do not focus on or produce narratives with much description of supervision and organizational aspects to the extent necessary to sufficiently populate these categories in either technique. The HERA technique captures supervision and organization causal factors under the Terms of Task, PSFs, and Information and Equipment. Of these Terms, only Supervision as a Task was used (N = 16, score = .17) in analysis of the incidents, further evidence that the incident reports did not have the information necessary to fully test these levels of either HFACS or HERA. Thus, it is difficult to compare the techniques on these causal factors, although the HFACS Terms representing them were selected in Phase 1 and appeared in the ranking task in Phase 2 (Table 6).

With these considerations in mind, there are several possible interpretations of the relative scores in Table 4. One is that perhaps a more rigorous training process would have changed the results. Another is

that the expert forum found concepts describing controller's behavior to be most useful for their operational purposes. That would explain why the terms scoring highest in each technique focused on individual behaviors. If the expert forum preferred more information about processes proximal to the individual's performance, then a harmonized technique should allow more detailed analysis of these categories in ways meaningful to the investigators. If this were true, in addition to the relative lack of the information in the narratives about supervisory activities and organizational influences, this might explain why the concepts describing supervision and the organization were ranked as less important to an understanding of the incident.

An alternate interpretation is that the expert forum's rankings were somehow reflective of differences between the techniques' lexicons, thus biasing the scoring results towards one or the other. Although both techniques used psychological terminology, the HERA technique was specifically developed for the ATC environment, whereas the version of HFACS used for this exercise was developed to classify causal factors from aviation accidents. Thus it is possible that the labeling and definition of concepts used by each technique also influenced the experts' selection and ranking of concepts.

Overall Relative Strengths and Weaknesses of the Techniques

After both teams completed their analyses, participants were asked for their oral and written opinions about how useful each technique would be relative to the other, including their strengths/weaknesses, and usability. The point of this activity was to identify the relative strengths and weaknesses of each technique from the point of view of operational personnel who are the target users of a harmonized technique. Their opinions would be considered, should the harmonization activity be undertaken. The questions asked and a summary of the most important results sorted by frequency are in Tables 7 through 9. These results may have been influenced by the type of training the SMEs received, but these results will be used to determine what characteristics of each technique will be valuable to include in activities to attempt harmonization.

Phase 3—JANUS: A Harmonized Technique

The goal of the third phase was to examine results from Phase 1 and 2 to determine whether harmonization was a feasible goal. It was desirable for a

Table 7: "What are the good/well-liked aspects of these approaches?"

HERA	N	HFACS	N
<i>Comments</i>		<i>Comments</i>	
A very comprehensive and detailed approach	10	The process is simple to understand and quick to use	9
Questions and flow-charts are good	4	Less time needed for analysis	2
Provides specificity	4	It describes items well	2
Leads the analyst through the process	4	There is a distinction between error and violation	2
Considers all errors in an event equally	3	Adverse supervision is considered a variable	2
Leads you back if you go wrong	2	Easier to train someone in this method	1
Does not blame	1	Includes causal factors	1

Table 8: "What are the poor / disliked aspects of these approaches?"

HERA	N	HFACS	N
<i>Comments</i>		<i>Comments</i>	
Too much paper to go through	3	Oversimplification which could lead to wrong conclusions	8
The Internal Error Modes, Psychological Error Mechanisms and Performance Shaping Factors are quite complex without training	3	Misunderstanding the tiers/categories/sub-categories	4
Adverse supervision should not be a Performance Shaping Factor	2	Limited nature of error classification	3
Too much human factors jargon	2	References to the pilot environment	3
Too subjective	2	Academic wording not suitable	2
The causal categories are difficult to establish	1	Definitions are not clear and specific enough	2
The pro-forma should be redesigned	1	Too easy to be subjective	2
Overlooks non-compliance from the controller	1	No cross checking in the technique	2
		Technique seems incomplete	2

Table 9: "What would you like to see included (✓) or excluded (x) in future technique development?"

HERA	✓	x	HFACS	✓	x
Recording Form	5	0	Unsafe Acts categories	6	0
Task lists	6	0	Error categories	6	2
Information and Equipment lists	4	2	Violation categories	5	3
Cognitive Domain flow charts	7	1	Preconditions for Unsafe Acts	3	4
External Error Mode/Violation tables	7	1	Unsafe Supervision categories	5	3
Internal Error Mode flow charts	7	1	Organizational Influence categories	6	2
Psychological Error Mechanism flow charts	7	1			
Performance Shaping Factors tables	7	1			

harmonized technique to include the best aspects from both HERA and HFACS. To fulfill this goal, the co-principal investigators held a four-day meeting in Brussels to discuss the findings from the previous two phases of work. During this meeting, the harmonized technique emerged.

Considerations

It was clear from this work that HERA and HFACS were developed for dissimilar objectives in two different domains. The different initial objectives and development of the two techniques had led, quite naturally, to different methodologies. Neither is better or worse; they are simply different. Although both techniques seek to address similar human factors issues, the method for identifying the issues and the granularity of analysis are different between the two approaches. Their commonality, however, is that they both draw from some of the same foundational literature of cognition and human error, albeit to different degrees and to different ends. Also, both attempt to improve how human error is identified and analyzed in the aviation environment. The goal of this work was to harmonize these different threads.

There are several points that should be mentioned about the objectives of these two methods. First, HFACS was originally developed with data from the military flight environment, although it has since been extended as a data analysis tool for other organizations. The HERA technique was specifically designed for incident analysis in the ATC environment. Second, and perhaps more important, is that HFACS was designed to investigate the human error embedded in aspects of an incident/accident as an event set within a larger system, whereas the HERA technique concentrates most analytical effort specifically on the human error causal factors in the incident. Although HERA captures human factors issues (e.g., in the PSF category of Organisational Factors\Management), HFACS specially tries to capture those categories (i.e., in Unsafe Supervision by Planned Inappropriate Operations). Another factor that likely influenced the usability and acceptability of the two approaches is the precise nature of the HERA technique, which was designed to find the specific underlying cognitive failure within the human, the controller in this case. The categorical HFACS technique, on the other hand, seeks to establish the chain of events to link the system vulnerabilities that result in failed human performance. Harmonization attempts to capture both of these perspectives.

It is clear from the work to date that a harmonized technique would benefit from incorporating the HERA technique's detailed, comprehensive, complex, and more specific methodology at the individual level. This should lend increased precision to incident investigation. The expert forum participants reported their appreciation of HERA's logical and structured technique that reduces subjectivity. However, the analysts also reported that the relative complexity, and often specialized use of language, would make use of the technique more difficult for the users of this technique without special training.

Similarly, the harmonized technique would benefit from incorporating the system-wide approach from HFACS. Users reported that HFACS is a simple, easy-to-comprehend technique, which lacks the cognitive specificity of HERA but is comprehensive and defines contextual factors at the supervisory and organizational levels. Contextual factors are often found more distal from the final incident or accident but are often no less contributory. Its broader categorical approach to analysis allows quicker analyses and possibly less training to use effectively.

Another consideration when examining the results of the Phase 1 and Phase 2 activities should be the sufficiency of training on both techniques given to the SMEs involved in the comparisons and the number of cases used for practice of both techniques. The goal of training was to provide the SME participants with a level of insight about each technique so that they could provide feedback about how they viewed the strengths and weaknesses of each relative to their operational experiences and needs.

JANUS

JANUS is the harmonization of HERA and HFACS. The specifics about how harmonization was accomplished will be laid out in future reports. Only a general overview of the technique is presented here.

This project revealed that the two techniques, HERA and HFACS, were as complementary as they were different. Thus, the ability of the HFACS technique to capture supervisory and organizational vulnerabilities was combined with the specificity of the HERA technique to generate the harmonized technique called JANUS, named for the mythological guardian of citizens, who looked into both the past and the future, representing the philosophy of learning from past error situations in service of future aviation safety. The technique is diagnostic at the level of the

individual's cognitive processes but also views the individual as part of the larger human-computer-organizational system.

In the JANUS technique, the *human factor* can only be understood in terms of the Person performing a specific Task with a particular piece of Equipment in a specific Environment, which includes supervisory and organizational influences. The JANUS technique permits a closer and comprehensive look at why the event occurred. The specificity of HERA to identify cognitive processes was combined with the tier structure of HFACS. The resulting method has diagnostic capabilities at the individual level but also captures an extensive array of performance shaping factors which might be present in the situation, as well as supervisory and organizational influences. Relationships between factors at the system level can be linked to factors at the level of an individual's thought processes.

JANUS is structured to interpret each incident as a series of critical points where a human error influences the course of the event. These critical points occur over time and form links in the chain of events that finally result in the incident. This technique has several implications for analysis. Each critical point can be identified and each receive an in-depth analysis to identify associated system variables, and specific cognitive, behavioral, and system vulnerabilities. This enables causal factors to be identified at both the individual and system level. In turn, these data can be analyzed to generate meaningful information about both individual events and system trends, from "situation awareness" to organizational resource management.

Future work to develop the JANUS technique will examine it for usability and reliability. Lessons learned from the development of HERA suggest that, as the number of decisions required to classify an error point increase, the inter-rater agreement between users may decrease. For example, as users progressed through HERA's structure to increasingly specific causal factors, Kappa, an indicator of the level of inter-rater agreement, decreased. Similarly, early work to develop HFACS required some category refinements to increase inter-rater agreement.

To assess the reliability of the JANUS technique, several factors need to be considered. Certainly for operational use, one would need to be confident that similar incidents having similar causal factors would, in fact, produce similar causal factors from the JANUS analysis. However, the usual test of inter-rater agreement uses two (or more) raters to analyze the same incident record, and an index of inter-rater

agreement is then calculated. Therefore, the first issue is whether "first-order" or "second-order" data are used to test inter-rater agreement. If the JANUS technique is used by the persons who were involved in the incident (i.e., controlling the air traffic at the time), as part of the initial incident investigation to identify causal factors ("first-order" analysis), the issue of inter-rater agreement will probably need to be approached differently than if the JANUS technique is applied to analyze materials, such as incident reports, which are output from the investigation ("second-order" analysis).

Replication of a "first order" analysis for the purpose of testing inter-analyst agreement presents different problems, compared with replication of "second order" analysis. In first-order analysis, inter-rater agreement would be calculated based on having the person who was involved in the incident, and who performed the JANUS analysis the first time, run through the interview process a second time after a period of time had passed, and then reliabilities could be calculated. This method, even supported by incident re-creation as a memory aid, is confounded with the analyst's deliberation over time about the incident and, possibly, reconstruction and hindsight bias. In second-order analysis, the materials from the incident have already been compiled, analyzed, and summarized. Investigators may have already identified a list of causal factors. These reports can potentially include bias from the incident investigators' preconceived points of view regarding the incidents.

Related to this, another consideration for assessment of inter-rater agreement is the specification of exactly what should be analyzed. In second-order analysis, several things can affect inter-rater agreement, such as a) whether all information about the incident has been included in the materials, b) what points (i.e., errors or causal factors) are identified for inclusion in the analysis, and c) whether predetermined conclusions have been reached and have been included in the materials.

The core JANUS technique, by employing specific diagnostic questions, does not rely on "traditional" methods of incident analysis. Instead, the technique uses HERA's structured interview approach to fill in and expand upon the elements of the HFACS taxonomy so that all dimensions are covered. The questions are written for the ATC domain and are capable of capturing incident causal factors ranging from the level of the individual's cognitive mechanisms, the task being conducted, interactions with equipment,

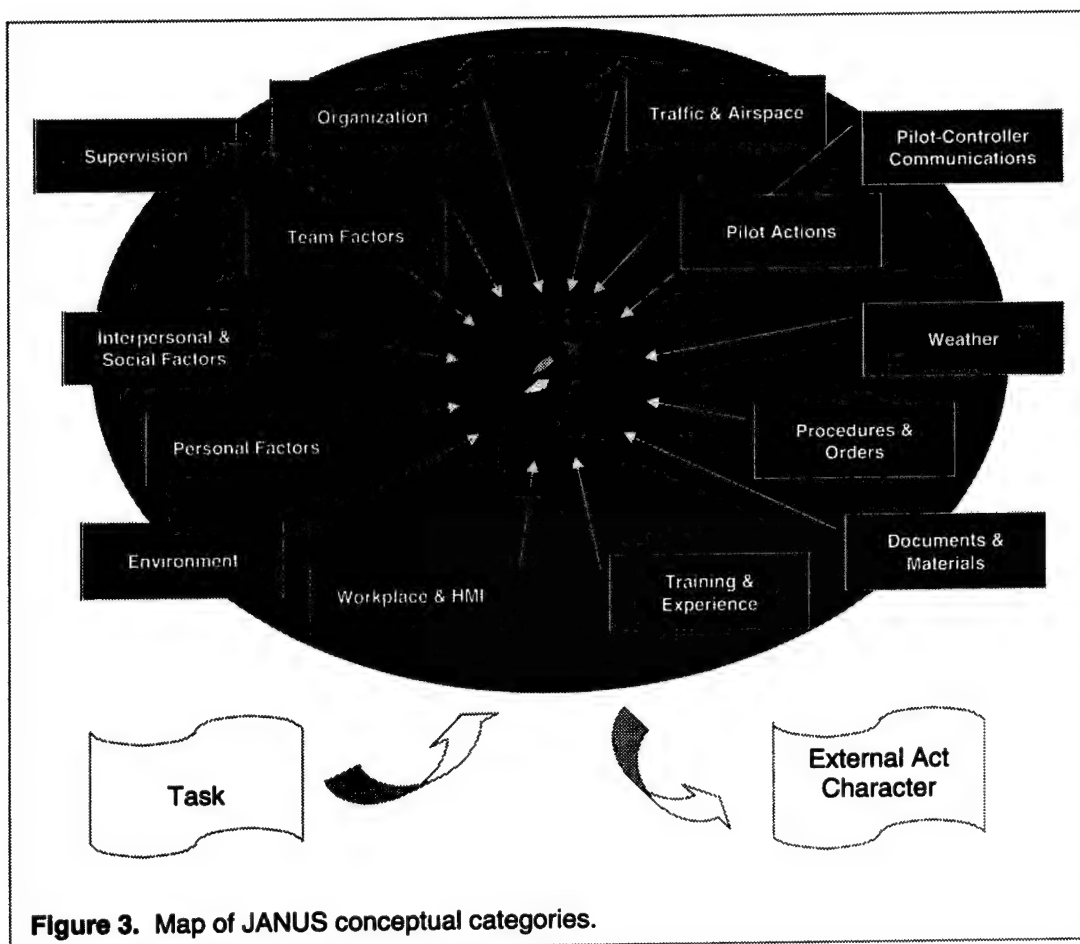


Figure 3. Map of JANUS conceptual categories.

to contextual conditions that shape performance (weather, traffic load, and equipment). The technique may prove useful for either first-order or second-order analysis, subject to its usability and reliability. Figure 3 shows the general conceptual groupings covered by this technique.

This comprehensive technique permits identification of causal and contributory factors so that appropriate "fixes" can be developed and applied. The problem with existing data collection techniques is that *data* does not equal *information*. Many organizations expend a lot of effort and resources gathering and archiving data, and publishing reports based on these data. However, mining such data does not often lend itself to generating meaningful information that can be turned into improving training programs, equipment development, or other remediation. For example, merely counting and reporting the number, type, and location of runway incursion events does not enable the development of effective mitigation strategies. Both HFACS and HERA have been developed to address this information problem.

CONCLUSION

This report compares each technique based on materials available at the time Action Plan 12 was initiated. Since then, each approach has continued to be developed separately by a variety of user groups interested in safety initiatives.

Prior to Action Plan 12, both HFACS and HERA were being developed as tools for post hoc analysis of incident reports written by the incident and accident investigators. In addition, the HFACS framework has been extended to other organizations and, in addition, has been used as an awareness aid for accident prevention.

Based on the input from the various SMEs involved in this project, the original goal of Action Plan 12 activities to harmonize two techniques into a single data mining tool has led quite naturally to discussions of whether the harmonized technique would be useful as an operational tool for ATC incident investigations staff to use for collecting first-order causal factors data. Should the harmonized technique eventually be

mature enough to be operationally deployed, then the tool should be able to be integrated with current quality assurance training and procedures for investigators. Not surprisingly, several development activities lay between the current point and that end state, including validating the usability and informativeness of the technique for operational use, assessment of the appropriate training for users of the technique, and mitigating any added operational workload.

JANUS is now undergoing an experimental trial in the US ATM system and by seven member States in the European Civil Aviation Conference. It is being tested as a tool to increase the information about causal factors related to operational errors. JANUS will be tested with operational error incidents by investigators and human factors researchers. At the completion of this testing phase, the technique will be reviewed and evaluated for its validity and utility as an investigatory tool.

As new systems are developed for ATM to meet future capacity demands, it is critical to have an understanding of the points at which human and system error might affect outcomes. It is likely that these tools will place increasing demands on the controller's cognitive processes to safely expedite air traffic. In addition to the known set of possible types of errors, new strategic planning tools are likely to introduce new types of errors. Once validated, the JANUS technique may provide a more sensitive means to identify and assess human and ATM system errors.

REFERENCES

- Bird, F. (1974). *Management guide to loss control*. Atlanta, GA: Institute Press.
- Dekker, S. (1999). *The re-invention of human error*. Paper presented at the 3rd International Human Error, Safety, and System Development Workshop, Liege, Belgium.
- Edwards, E. (1988). Introductory overview. In E. L. Wiener & D. C. Nagel (Eds.), *Human factors in aviation* (pp. 3-25). San Diego: Academic.
- EATMP. (1999a). Short Report on Human Performance Models and Taxonomies of Human Error in Air Traffic Management. (HUM.ET1.ST). Brussels: EUROCONTROL.
- EATMP. (1999b). Technical review of human performance models and taxonomies of human error in air traffic management. (HUM.ET1.ST). Brussels: EUROCONTROL.
- EATMP. (2000). *Validation of the human error in ATM (HERA) classification system*. Draft report. Brussels: EUROCONTROL.
- Federal Aviation Administration. (2001). *Air traffic quality assurance* (FAA Order 7210.56). Washington, DC: Department of Transportation.
- Fleiss, J.L. (1981). *Statistical methods for rates and proportions*. New York: Wiley.
- Heinrich, H., Petersen, D., & Roos, N. (1931). *Industrial accident prevention: A safety management approach* (1st ed.). New York: McGraw-Hill.
- Hollnagel, E., Cacciabue, P.C., & Hoc, J. (1995). Work with technology: Some fundamental issues. In J. Hoc, P.C. Cacciabue, and E. Hollnagel (Eds.), *Expertise and technology* (pp. 1-15). Hillsdale, NJ: Erlbaum.
- Isaac, A.R., & Ruitenbergh, B. (1999). *Air traffic control: Human performance factors*. Hants, England: Ashgate.
- Kinney, G.C., Spahn, M.J. & Amato, R.A. (1977). The human element in air traffic control: Observations and analyses of the performance of controllers and supervisors in providing ATC separation services (MTR-7655). McLean, VA: MITRE Corporation, METRIEK Division.
- Jensen, R. (1995). *Pilot judgment and crew resource management*. Brookfield, VT: Avebury Aviation.
- Martiniuk, R.G. (1976). *Information processing in motor skills*. New York: Rinehart and Winston.
- O'Hare, D., Wiggins, M., Batt, R., & Morrison, D. (1994). Cognitive failure analysis for aircraft accident investigation. *Ergonomics*, 37(11), 1855-69.
- Rasmussen, J. (1982). Human errors: A taxonomy for describing human malfunction in industrial installations. *Journal of Occupational Accidents*, 4, 311-33.
- Reason, J. (1990). *Human error*. New York: Cambridge University Press.

- Redding, R.E. (1992). Analysis of operational errors and workload in air traffic control. *Proceedings of the Human Factors Society 36th Annual Meeting* (pp. 1321-5), Vol. 2. Santa Monica, CA: Human Factors Society.
- Rodgers, M.D., & Nye, L.G. (1993). Factors associated with the severity of operational errors at air route traffic control centers. In M.D. Rodgers (Ed.), *An examination of the operational error database for air route traffic control centers* (pp. 11-25). Washington, DC: Federal Aviation Administration (Report No. DOT/FAA/AM-93/22), ADA 275986.
- Schroeder, D.J. (1982). The loss of prescribed separation between aircraft: How does it occur? *Proceedings (P-114), Behavioral Objectives in Aviation Automated Systems Symposium*, Society of Automotive Engineers (SAE), 257-69.
- Schroeder, D.J., & Nye, L.G. (1993). An examination of the workload conditions associated with operational errors/deviations in air route traffic control centers. In M.D. Rodgers (Ed.), *An examination of the operational error database for air route traffic control centers* (pp. 1-10). Washington, DC: Federal Aviation Administration (Report No. DOT/FAA/AM-93/22), ADA 275986.
- Shappell, S.A., & Wiegmann, D.A. (1997). A human error approach to accident investigation: The Taxonomy of Unsafe Operations. *International Journal of Aviation Psychology*, 7(4), 269-91.
- Shappell, S.A., & Wiegmann, D.A. (2000). The Human Factors Analysis and Classification System—HFACS. (DOT/FAA/AAM-00/07). Washington, DC: Federal Aviation Administration.
- Shappell S., & Wiegmann, D. (2001). Uncovering human error trends in fatal general aviation accidents using HFACS. Presented at the 72nd Annual Meeting of the Aerospace Medical Association, Reno, NV.
- Stager, P., & Hameluck, D. (1990). Ergonomics in air traffic control. *Ergonomics*, 33, 493-9.
- Wickens, C.D. (1992) *Engineering Psychology and Human Performance* (Second ed.). New York: Harper-Collins.
- Wickens, C.D., & Flach, J.M. (1988). Information processing. In E.L. Wiener & D.C. Nagel (Eds.), *Human factors in aviation* (pp. 111-55). San Diego: Academic.
- Wiegmann, D.A., & Shappell, S.A. (1997). Human factors analysis of postaccident data: Applying theoretical taxonomies of human error. *International Journal of Aviation Psychology*, 7(1), 67-81.
- Wiegmann, D., & Shappell, S. (2001a). *Assessing the reliability of HFACS within the context of aviation*. Presented at the 72nd Annual Meeting of the Aerospace Medical Association, Reno, NV.
- Wiegmann, D.A., & Shappell, S.A. (2001b). Human error analysis of commercial aviation accidents: Application of the Human Factors Analysis and Classification System (HFACS). *Aviation, Space, and Environmental Medicine*, 72(11), 1006-16.

APPENDIX A

A Summary of the Comparisons (See text for citations)

	HERA	HFACS
<i>Origin</i>	Developed for incident analysis of human errors in ATM	Developed for incident analysis of human errors in aviation accidents
<i>Theoretical Base</i>	Human Error Taxonomies Human Performance Models Task-based Taxonomies Error Mode Taxonomies Communication System Models Information Processing Models Symbolic Processing Cognitive Simulations Other Models and Taxonomies, e.g., SA, control system, SDT, commission errors approaches, violations Other Domain Approaches, e.g., accident theory, root cause analysis, nuclear risk assessment, maritime operations, flight operations, ATM Models of Error in ATM performance	Human Error Taxonomies Human Performance Models Industrial Safety Information Processing Models Crew Resource Management
<i>Conceptual Coverage</i>	Ranges from the organizational level to individual internal psychological mechanisms (e.g., expectation bias).	Ranges from the organizational level to the individual's error (i.e., decision, skill, misperception).
<i>Data for Analysis</i>	Incident report data and narrative summaries.	Lists of causal factors from incident databases in the context of the narrative summaries.
<i>Analytical Process</i>	Each human error point within the incident description is subjected to the entire HERA analysis. Error points are identified by working from the beginning of the incident report to the final event. Analyst is led through the technique to the causal factors via a structured query process. Analysis is a re-analysis of the incident	The Unsafe Act is identified as well as each related classifiable act in the incident description. Each is then categorized by working backwards from the Unsafe Act. Classifiable acts are identified as "holes in the cheese." Analyst assigns each given causal factor to the appropriate cells in the taxonomy. Analysis is not a re-analysis of the incident.

(Continued)

APPENDIX A (continued)

A Summary of the Comparisons (See text for citations)

<i>Inter-coder Reliability</i> Values of $k = .40$ or less are considered "poor" agreement while values of $k = .75$ or greater are considered "excellent" levels of agreement (Fleiss, 1981).	At the level of Cognitive Domains, Kappa ranged from .44 to .50. With extended training, coders showed overall increased agreement (Kappa = .52), compared to .38 with only basic training. By job function the incident investigators showed highest agreement (Kappa = .61). ATM and researchers agreement was .23 and .43 respectively. Agreement between coders declined as the level of analysis becomes finer-grained, although psychological specificity increases.	Pair-wise comparisons of inter-rater agreement using Cohen's Kappa ranged from .60 in early studies to .95 later in development of the model. Using all categories, Kappa ranged from .65 to .75. Reported inter-rater agreement was lowest for the Supervisory and Organizational tiers.
<i>Output Data</i>	Each human error can be described by a cognitive domain, internal error mode, and psychological error mechanism. Each error can also be identified by the associated task, information, and a variety of situational performance shaping factors.	Each classified act can be labeled by HFACS tier, category within tier, and subcategory within category if available. Each data point has an associated description which can be subjected to content analysis.